Phân Tích Dữ Liệu Thực Tế với Python

Bài 8.2: Phân Tích EDA Cơ Bản

Quang-Khai Tran, Ph.D
CyberLab, 03/2023

(Ảnh: Internet)

CYBERLEARN
ĐÀO TẠO CHUYÊN GIA LẬP TRÌNH

CYBERSOFT
ĐÀO TẠO CHUYÊN GIA LẬP TRÌNH

CYBERLAB
ĐÀO TẠO CHUYÊN GIA LẬP TRÌNH

Nội dung

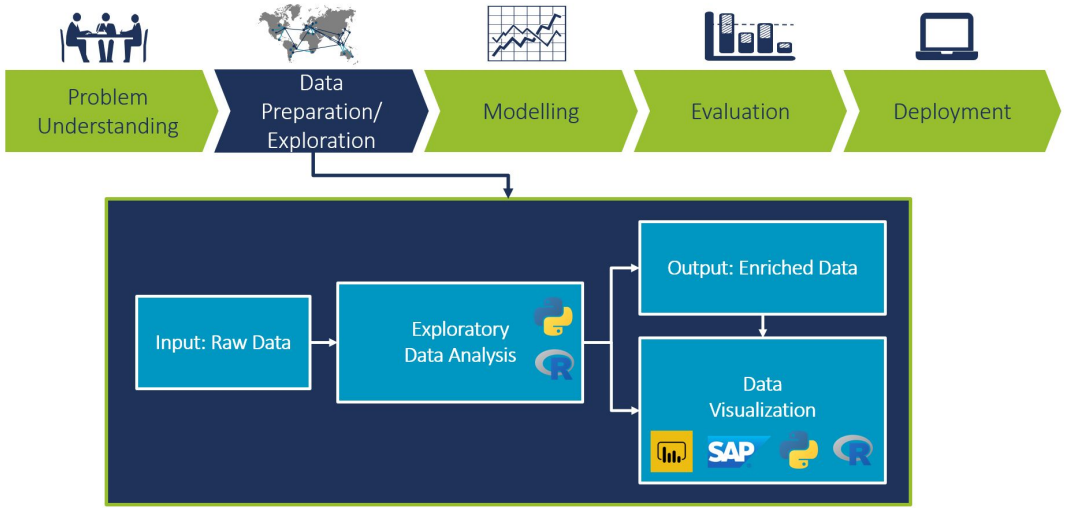
1. Giới thiệu
2. Một số phân tích EDA cơ bản
3. Demo: Dữ liệu InstaCart Market Basket
4. Bài tập & Thảo Luận

Phần 1. Giới thiệu

Exploratory Data Analysis (EDA)

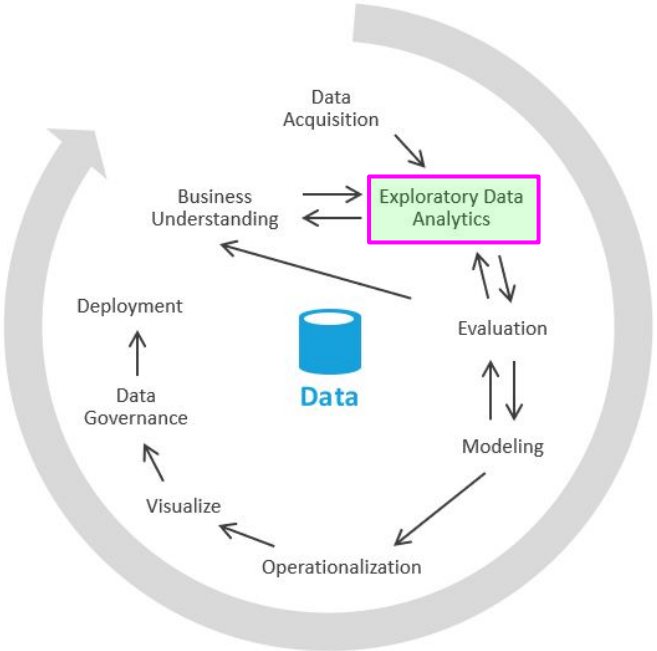
Phân tích EDA (khai phá dữ liệu) là gì?

- Là bước phân tích cơ bản được thực hiện sớm khi tiến hành phân tích dữ liệu
- Thường sử dụng các biểu đồ



Tham khảo: https://blog.camelot-group.com/2019/03/exploratory-data-analysis-an-important-step-in-data-science/

Phân tích EDA (khai phá dữ liệu) là gì?



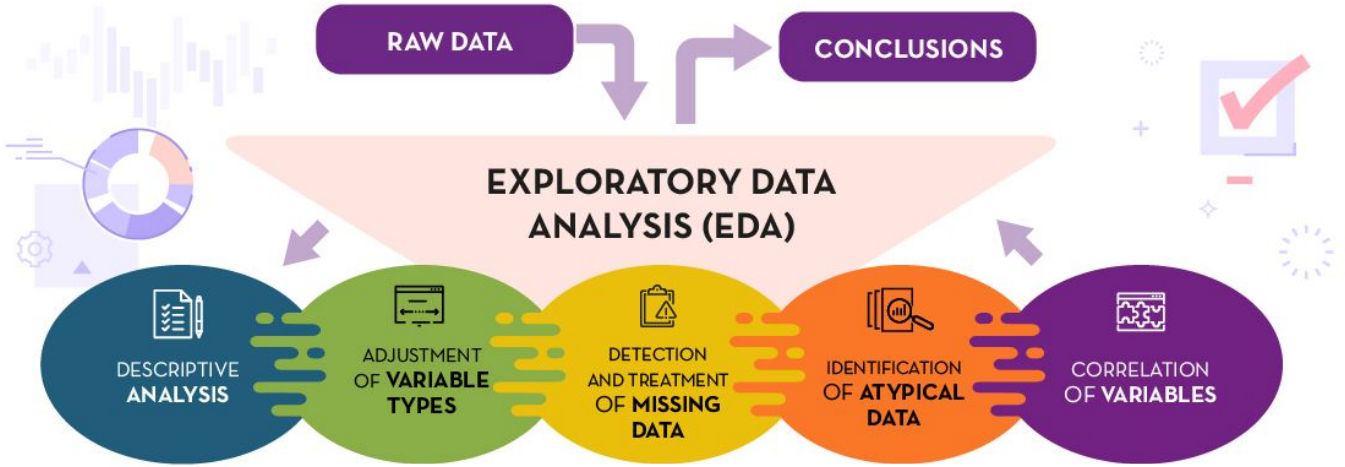
Source: Exploratory data analysis with Azure Synapse... | Microsoft

- Có thể thực hiện luôn với:
  - Dữ liệu thô
  - Chưa làm sạch

1 Giới thiệu

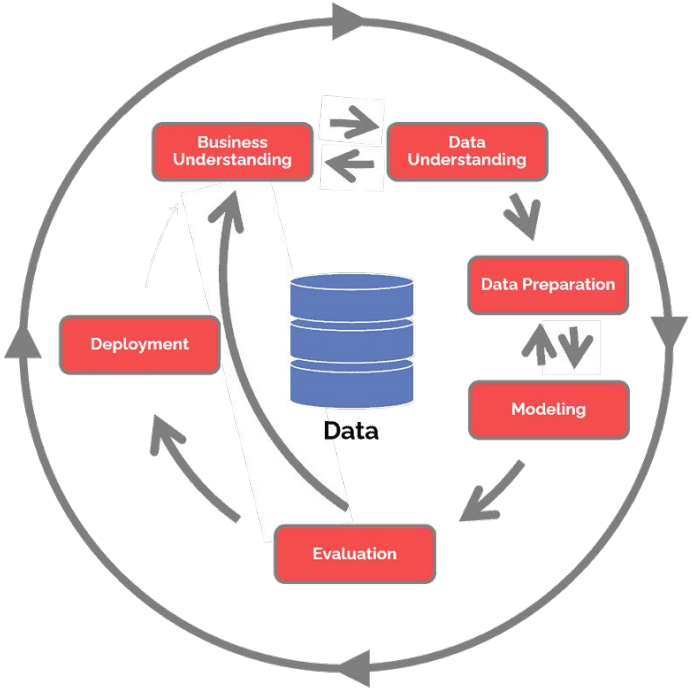
- Chủ yếu tóm tắt các đặc điểm/thuộc tính chính trên dữ liệu:
  - Kích thước, độ lớn của tập dữ liệu
  - Ý nghĩa, kiểu dữ liệu của từng trường dữ liệu
  - Các đặc điểm/thuộc tính của từng trường dữ liệu
  - Thống kê các dữ liệu khuyết
  - Tìm các outliers
  - Tìm các dòng (record) hay giá trị bị lặp lại
  - Các đặc điểm/thuộc tính giữa các trường dữ liệu với nhau: nguyên nhân (causes) và quan hệ (relationships)

Phân tích EDA (khai phá dữ liệu) là gì?



Source: A Practical Introductory Guide to Exploratory Data Analysis

❖ EDA in CRISP-DM

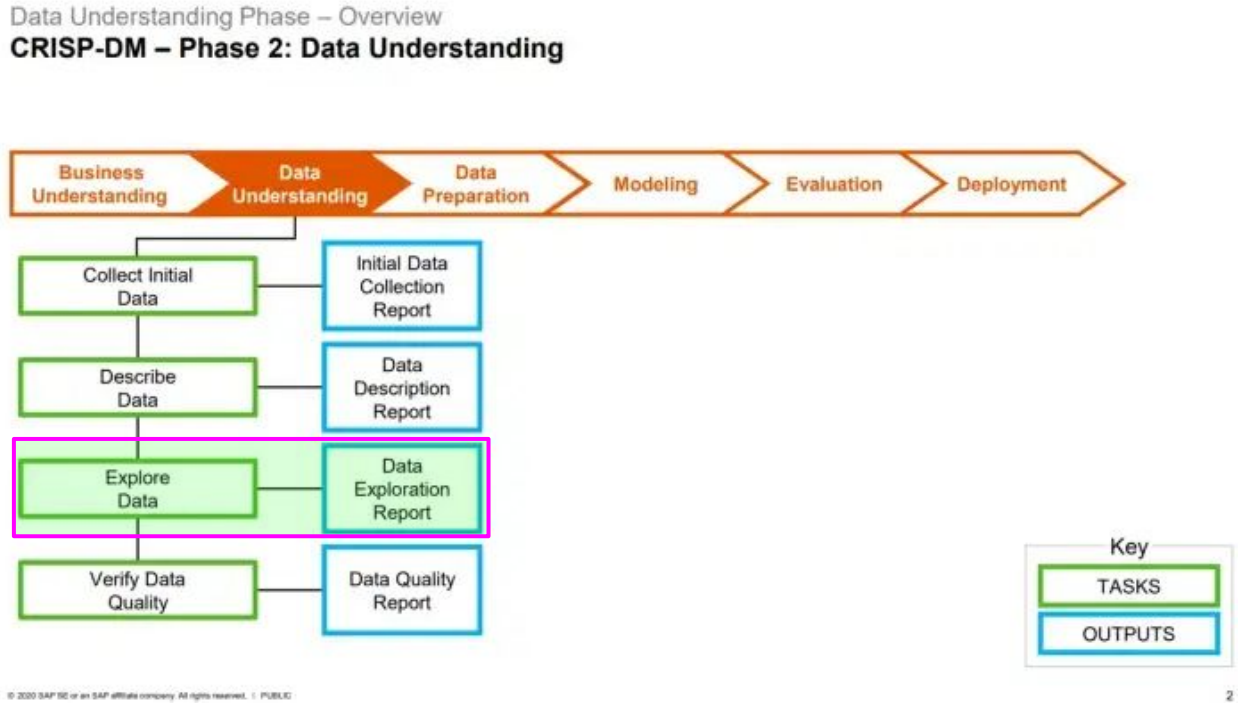


CRoss Industry Standard Process for Data Mining (CRISP-DM)

1. Business understanding – What does the business need?
2. Data understanding – What data do we have / need? Is it clean?
3. Data preparation – How do we organize the data for modeling?
4. Modeling – What modeling techniques should we apply?
5. Evaluation – Which model best meets the business objectives?
6. Deployment – How do stakeholders access the results?

Source: https://www.datascience-pm.com/crisp-dm-2/

❖ EDA in CRISP-DM (Cross Industry Standard Process for Data Mining)



Mục đích chính của EDA

“Giúp chúng ta có một cái nhìn khởi đầu vào dữ liệu trước khi đưa ra bất cứ giả định nào” (IBM)

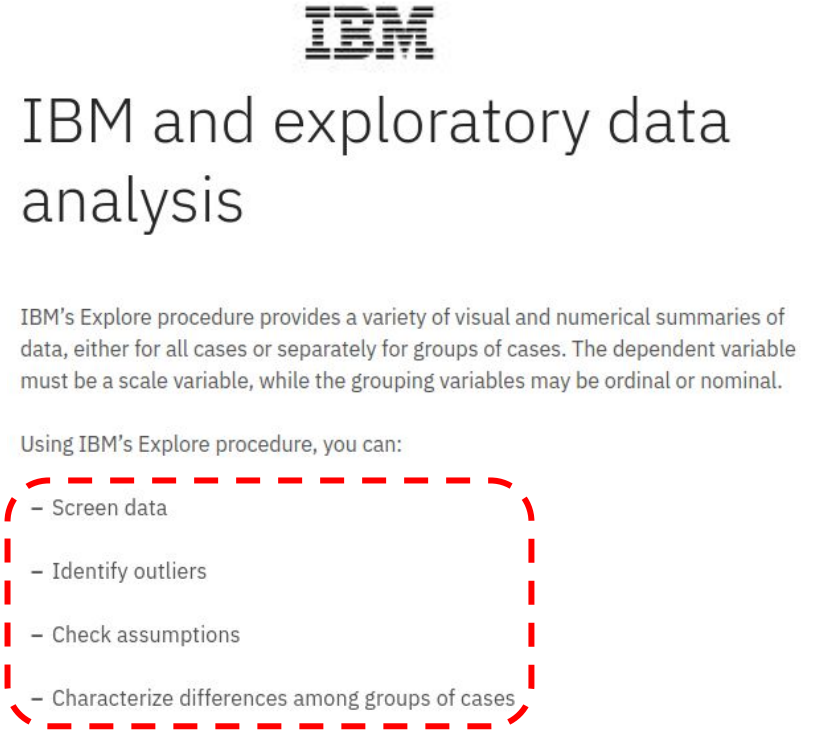
Phân tích EDA là gì?

IBM

1. Phân tích đơn biến không biểu đồ	- Là dạng phân tích mô tả trên một biến - Bao gồm các số liệu của phân tích mô tả
2. Phân tích đơn biến với biểu đồ	- Cũng là phân tích mô tả trên một biến - Bao gồm các biểu đồ của phân tích mô tả
3. Phân tích đa biến không biểu đồ	- Là dạng phân tích mô tả trên mối quan hệ của hai hay nhiều biến và không dùng biểu đồ
4. Phân tích đa biến với biểu đồ	- Là dạng phân tích mô tả trên mối quan hệ của hai hay nhiều biến dùng biểu đồ

Ví dụ về quy trình phân tích EDA của IBM

1. "Đọc" dữ liệu
2. Xác định các giá trị ngoại biên
3. Kiểm tra các giả định
4. Chỉ ra các đặc trưng khác nhau giữa các nhóm trường hợp/vấn đề



IBM

IBM and exploratory data analysis

IBM's Explore procedure provides a variety of visual and numerical summaries of data, either for all cases or separately for groups of cases. The dependent variable must be a scale variable, while the grouping variables may be ordinal or nominal.

Using IBM's Explore procedure, you can:

- Screen data
- Identify outliers
- Check assumptions
- Characterize differences among groups of cases

Lợi ích của Phân tích EDA?

- Giúp nắm bắt được các đặc trưng cơ bản của dữ liệu
  - Kích thước, độ phức tạp
  - Chất lượng của dữ liệu, những thiếu sót, sai sót
  - Tiềm năng của dữ liệu
- Đưa ra được các hướng phân tích tiếp theo
- Giúp lựa chọn công cụ, phương pháp phân tích phù hợp cho các hướng đó

…

Các cột dữ liệu trong bảng

- Trước khi xử lý, làm sạch: data fields (trường dữ liệu/trường thông tin)
- Sau khi xử lý, làm sạch: data features (đặc trưng/thuộc tính)

Chu trình cơ bản của phân tích EDA

1. Đưa ra các câu hỏi/yêu cầu về dữ liệu
2. Tìm câu trả lời thông qua việc chuyển đổi dữ liệu, làm sạch, thực hiện một số phân tích mô tả, vẽ biểu đồ và lập mô hình dữ liệu.
3. Sử dụng các kết quả để cải thiện các câu hỏi cũ, tạo ra câu hỏi mới.

Phân tích EDA là một chu trình lặp

Tuy nhiên phân tích EDA chưa được chuẩn hóa với các quy tắc thống nhất, bạn có thể tự sáng tạo các metrics, cách nhìn mới

Tham khảo: https://r4ds.had.co.nz/exploratory-data-analysis.html

1 Phân Nhóm Các Loại dữ liệu trong thống kê

Quantitative (định lượng):

- discrete (rời rạc)
- continuous (liên tục)
- interval (khoảng)

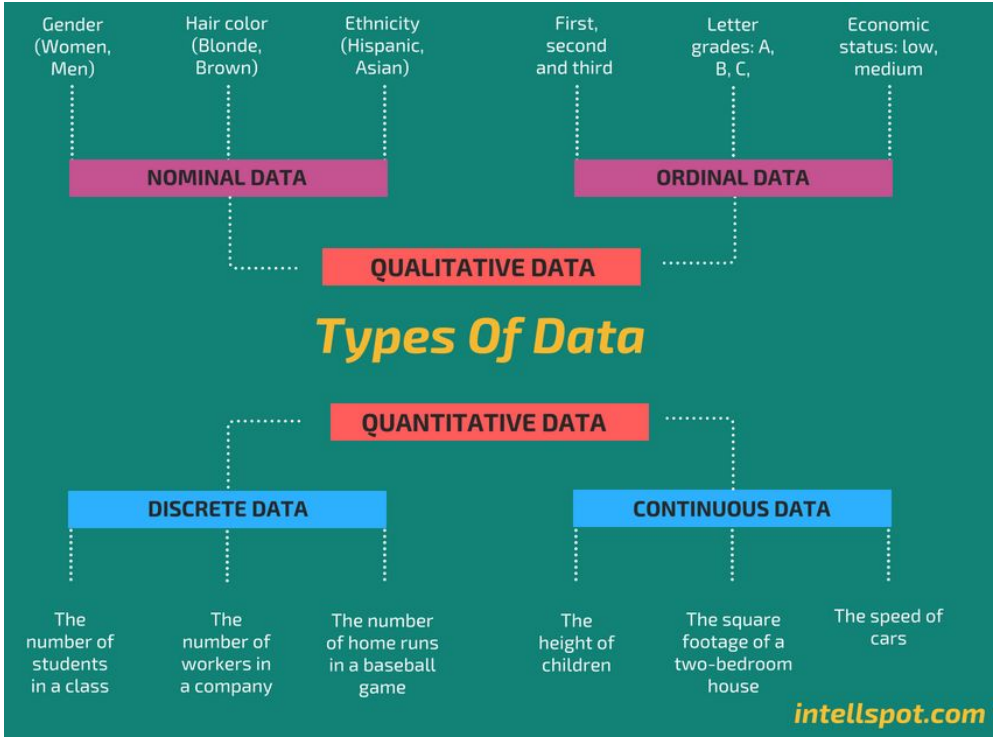
  ⇒ structured data

Qualitative (định tính):

- nominal (định danh)
- binary (định danh True/False)
- ordinal (thứ tự)

  ⇒ unstructured data

(text, phân loại, datetime)



Source: https://www.intellspot.com/data-types/

Phần 2. Một số phân tích EDA cơ bản

2.1. Phân tích đơn biến
2.2. Phân tích đa biến

2.1 Phân tích đơn biến

Các yếu tố cần xem xét

1. Các trường dữ liệu	Tên, ý nghĩa, loại dữ liệu, số dòng bị sai (error), thiếu (missing data)
2. Độ đo về tần số/tần suất	Count/Percentage
3. Độ đo về khuynh hướng tập trung	Mean, Median, Mode
4. Độ đo về sự biến thiên/mở rộng	Min - Max, Range, Variance, Std
5. Độ đo về phân vị	Percentile, Quantile, Quartile ⇒ tìm outliers
6. Độ đo về dạng phân bố	Normal distribution, Skewness, Kurtosis

- Kích thước dữ liệu: nếu quá nhỏ?
- Tên và ý nghĩa: một số trường dữ liệu bị nặc danh hóa và không thể biết
- Phân phối xác suất trong cột dữ liệu:
  - Mọi giá trị trong cột bằng nhau, hoặc chỉ xuất hiện vài giá trị:
    không có nhiều ý nghĩa, có thể xóa cột
  - Nếu có quá nhiều dữ liệu bị thiếu:
    có thể ảnh hưởng lớn đến kết quả, cần thận trọng và tìm cách phù hợp
  - Các giá trị không hợp lệ: cần coi là bị thiếu (ví dụ tuổi khách hàng < 0)
  - Xuất hiện outliers: cần có cách xử lý phù hợp

2.1 Phân tích đơn biến

Các phân tích thuộc tính/đặc điểm

- Các thông tin về sản phẩm/hàng hóa và phân loại của chúng
- Các thông tin về thời gian, khoảng thời gian, tính chu kỳ
- Các thông tin về không gian, địa điểm
- Các thông tin về người, nhân lực, nhân sự

…

Ví dụ

- Doanh số bán hàng theo loại mặt hàng, theo ngành hàng
- Doanh số bán hàng theo quận, huyện, thành phố, khu vực
- Doanh số bán hàng theo ngày, giờ, tuần, ngày trong tuần

…

2.2 Phân tích đa biến

1. Phân tích tương quan	Hệ số tương quan (R), hệ số xác định (R^2)
2. Phân tích hồi quy	Mô hình dự đoán dự vào linear regression và logistic regression (nếu có thể)

2.2 Phân tích đa biến

Ví dụ: quan hệ giữa trường dữ liệu cần dự đoán và các trường thông tin còn lại

- Nếu mối tương quan là yếu, trường thông tin có thể ít có ý nghĩa
- Nếu mối tương quan là mạnh, cần giữ lại và tập trung xử lý, làm sạch
- Nếu 2 trong số các trường còn lại có sự tương quan cao, cần kiểm tra lại xem có thể bỏ qua một cột hay không

2 Tham khảo thêm

Sách "Machine Learning cơ bản" (Vũ Hữu Tiệp)

- Bài EDA:
  https://machinelearningcoban.com/tabml_book/ch_data_processing/eda_purpose.html
- Demo với dữ liệu Titanic:
  https://machinelearningcoban.com/tabml_book/ch_data_processing/eda_titanic.html
- Demo với dữ liệu California Housing:
  https://machinelearningcoban.com/tabml_book/ch_data_processing/eda_cali_housing.html

Phần 3. Demo phân tích EDA cơ bản

Dữ liệu InstaCart Market Basket

3.1 Dữ liệu InstaCart Market Basket

3.1 Dữ liệu InstaCart Market Basket

Dữ liệu giỏ hàng:

- Dữ liệu về đơn hàng của khách hàng trên hệ thống, đã được nặc danh hóa
- Sử dụng cho một cuộc thi trên Kaggle (2017)
- Gồm hơn 3.4 triệu orders và nhiều thông tin khác liên quan
- Được sử dụng rất phổ biến trong các tutorial về DA/DS

fresh fruits	3642188
fresh vegetables	3418021
packaged vegetables fruits	1765313
yogurt	1452343
packaged cheese	979763
milk	891015
water seltzer sparkling water	841533
chips pretzels	722470
soy lactosefree	638253
bread	584834

- 3,421,083 đơn hàng (orders)
- 49,688 loại sản phẩm (products)
- 134 loại kệ hàng (aisles)
- 21 loại gian hàng (departments)

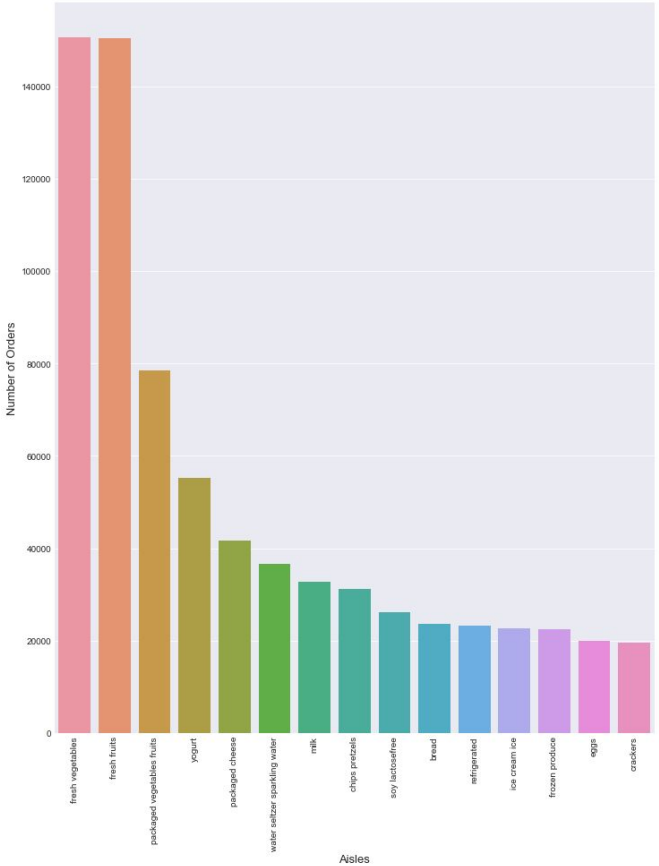
3.1 Dữ liệu InstaCart Market Basket

Dữ liệu giỏ hàng:

- Dữ liệu về đơn hàng của khách hàng trên hệ thống, đã được nặc danh hóa
- Sử dụng cho một cuộc thi trên Kaggle
- Gồm hơn 3 triệu orders và nhiều thông tin khác liên quan
- Được sử dụng rất phổ biến trong các tutorial về DA/DS

Link: https://www.kaggle.com/c/instacart-market-basket-analysis

3.1 Dữ liệu InstaCart Market Basket



3.2 Demo: Phân tích đơn hàng từ 1 file

File 'orders.csv', có 7 fields:

1. 'order_id' --> id của đơn hàng
2. 'user_id', --> id của khách hàng
3. 'eval_set' --> biểu thị datasets mà dòng này thuộc về
   (dùng trong competition của nhà tổ chức, ở đây ta không cần quan tâm)
4. 'order_number' --> thứ tự của order trong các lần order của một khách hàng
5. 'order_dow' --> ngày order trong tuần (date of week)
6. 'order_hour_of_day' --> giờ order trong ngày (0-23)
7. 'days_since_prior_order' --> số ngày kể từ lần order trước đó của một khách hàng

Link: https://www.kaggle.com/c/instacart-market-basket-analysis

3.3 Demo: Phân tích đơn hàng từ nhiều files

File 'order_products__train.csv', có 4 fields:

1. 'order_id' --> id của đơn hàng
2. 'product_id', --> id của sản phẩm được ordered
3. 'add_to_cart_order' --> thứ tự mà sản phẩm được đưa vào order
4. 'reordered' --> sản phẩm có phải là được reordered bởi khách hàng này hay không

File 'products.csv', có 4 fields:

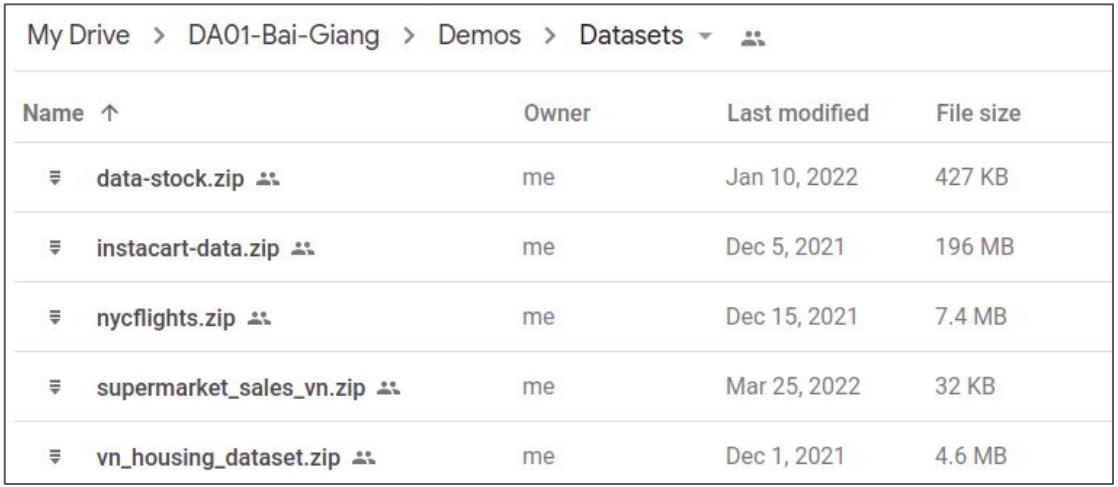
1. 'product_id', --> id của sản phẩm
2. 'product_name' --> tên của sản phẩm
3. 'aisle_id' --> id của kệ hàng mà sản phẩm thuộc về
4. 'department_id' --> id của gian hàng mà sản phẩm thuộc về

Download: https://drive.google.com/drive/folders/1xKXFshk1DxBQKXcX_hJGCyb2NZFATzIn?usp=sharing

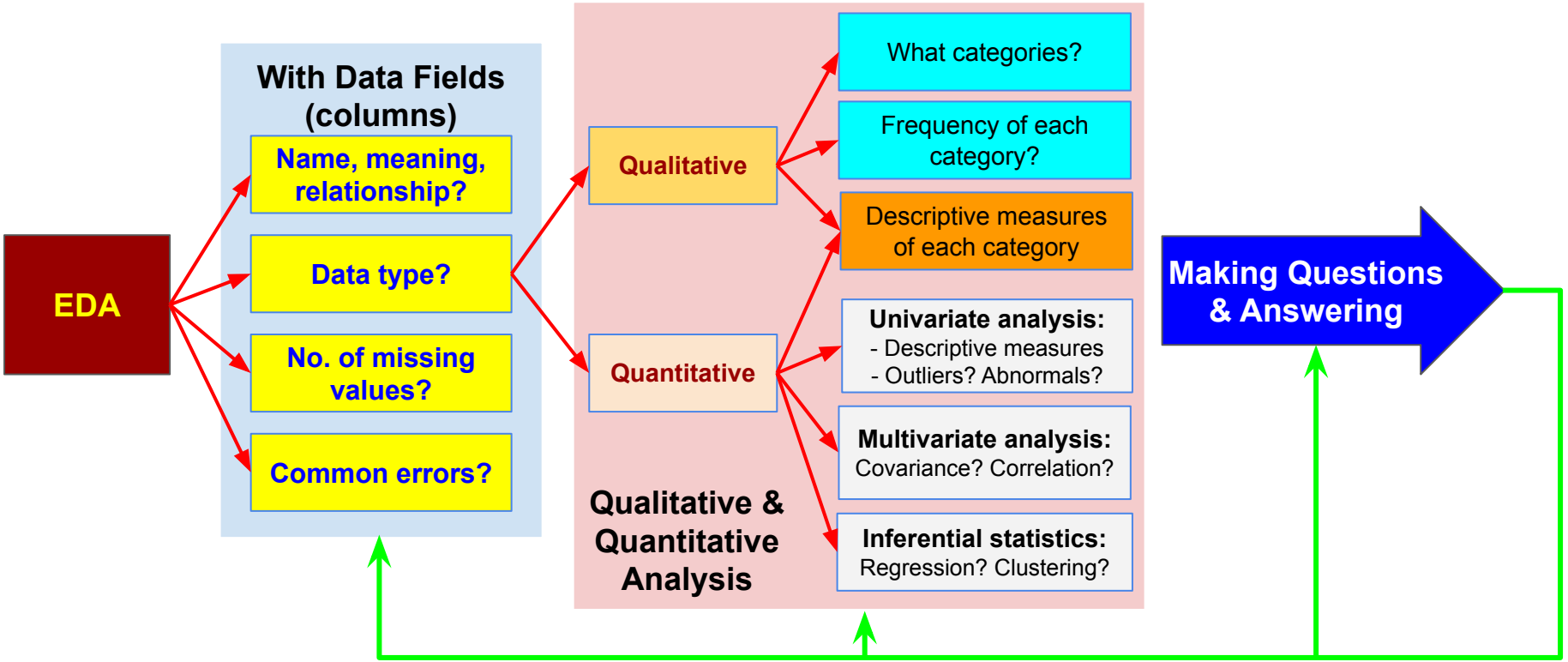
3.3 Demo: Phân tích đơn hàng từ nhiều files

Download:

https://drive.google.com/drive/folders/1l73pdoNfKWludLhl2p9Cb9cSfJkNf2no?usp=sharing



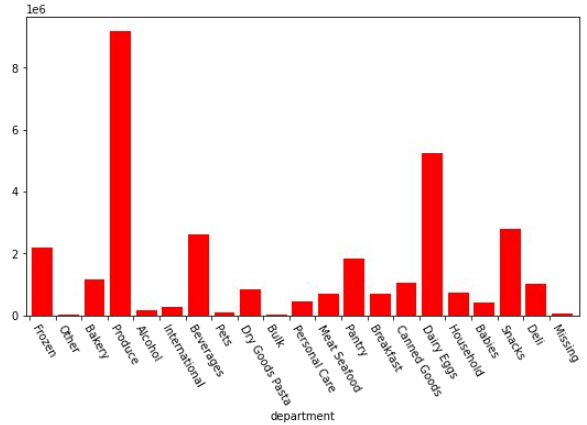
4 Tóm tắt phân tích EDA

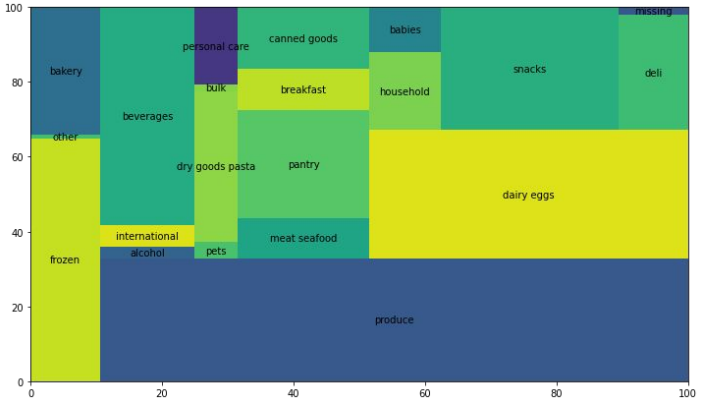


4 Bài tập

Phân tích EDA cho dữ liệu **product** trong 'order_products__prior.csv':

1. Tìm tên 10 sản phẩm bán được nhiều nhất và vẽ biểu đồ
2. Tìm tên 10 sản phẩm được re-ordered nhiều nhất và vẽ biểu đồ
3. Thống kê số lượng sản phẩm bán được của các department và vẽ biểu đồ
   - Biểu đồ cột
   - Biểu đồ tree-map





THANK YOU!